A robust phylogenomic timetree for biotechnologically and medically important fungi from Aspergillaceae (Eurotiomycetes, Ascomycota)

Jacob L. Steenwyk[1], Xing-Xing Shen[1], Abigail L. Lind[2,3], Gustavo H. Goldman[4], and Antonis Rokas[1,2,*]

[1] Department of Biological Sciences, Vanderbilt University, Nashville, TN, 37235, USA

[2] Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America

[3] Gladstone Institute for Data Science and Biotechnology, San Francisco, California, USA

[4] Departamento de Ciências Farmacêuticas, Faculdade de Ciências Farmacêuticas de Ribeirão Prêto, Bloco Q, Universidade de São Paulo, São Paulo, Brazil

* Materials and correspondence: antonis.rokas@vanderbilt.edu

Running title: Phylogenomics of Aspergillaceae

Keywords: Incongruence, genomics, phylogenetics, Eurotiomycetes

Abbreviations: NT, nucleotide; AA, amino acid; CI, confidence interval; RCV, relative composition variability; IC, internode certainty; GSF, gene support frequencies; GLS, gene-wise log-likelihood scores; DVMC, degree of violation of a molecular clock;

Abstract

The filamentous fungal family Aspergillaceae contains > 1,000 known species, mostly in the genera *Aspergillus* and *Penicillium*. Fungi in Aspergillaceae display a wide range of lifestyles, including several that are of relevance to human affairs. For example, several species are used as industrial workhorses, food fermenters, or platforms for drug discovery (e.g., *Aspergillus niger, Penicillium camemberti*), while others are dangerous human and plant pathogens (e.g., *Aspergillus fumigatus, Penicillium digitatum*). Reconstructing the phylogeny and timeline of the family's diversification is the first step toward understanding how its diverse range of lifestyles evolved. To infer a robust phylogeny for Aspergillaceae and pinpoint poorly resolved branches and their likely underlying contributors, we used 81 genomes spanning the diversity of *Aspergillus* and *Penicillium* to construct a 1,668-gene data matrix. Phylogenies of the nucleotide and amino acid versions of this full data matrix were generated using three different maximum likelihood schemes (i.e., gene-partitioned, unpartitioned, and coalescence). We also used the same three schemes to infer phylogenies from five additional 834-gene data matrices constructed by subsampling the top 50% of genes according to different criteria associated with strong phylogenetic signal (alignment length, average bootstrap value, taxon completeness, treeness / relative composition variability, and number of variable sites). Examination of the topological agreement among these 36 phylogenies and measures of internode certainty identified 12 / 78 (15.4%) bipartitions that were incongruent. Patterns of incongruence across these 12 bipartitions fell into three categories: (i) low levels of incongruence for 2 shallow bipartitions, most likely stemming from incomplete lineage sorting, (ii) high levels of incongruence for 3 shallow bipartitions, most likely stemming from hybridization or introgression (or very high levels of incomplete lineage sorting), and (iii) varying levels of incongruence for 7 deeper bipartitions,

45 most likely stemming from reconstruction artifacts associated with poor taxon sampling. Relaxed
46 molecular clock analyses suggest that Aspergillaceae likely originated in the lower Cretaceous,
47 125.1 (95% Confidence Interval (CI): 146.7 - 102.1) million years ago (mya), with the origins of
48 the *Aspergillus* and *Penicillium* genera dating back to 84.3 mya (95% CI: 90.9 - 77.6) and 77.4
49 mya (95% CI: 94.0 - 61.0), respectively. Our results provide a robust evolutionary and temporal
50 framework for comparative genomic analyses in Aspergillaceae, while our general approach
51 provides a widely applicable template for phylogenomic identification of resolved and
52 contentious branches in densely genome-sequenced lineages across the tree of life.

The vast majority of the 1,062 described species from the family Aspergillaceae (phylum Ascomycota, class Eurotiomycetes, order Eurotiales)[1] belong to the genera *Aspergillus* (42.5%; 451 / 1,062) and *Penicillium* (51.6%; 549 / 1,062)[2,3]. Fungi from Aspergillaceae exhibit diverse ecologies; for example, *Penicillium verrucosum* is widespread in cold climates but has yet to be isolated in the tropics[4], whereas *Aspergillus nidulans* is able to grow at a wide range of temperatures but favors warmer temperatures[5]. Several representative species in the family are exploited by humans, while a number of others are harmful to humans or their activities[6]. For example, *Aspergillus oryzae* is used in the production of traditional Japanese foods including soy sauce, sake, and vinegar[7,8], *Penicillium camemberti* and *Penicillium roqueforti* contribute to cheese production[9,10], *Aspergillus niger* is used in the production of enzymes that are later used in starch processing, baking and brewing industries, in animal feed, and the paper industry[11], and *Penicillium citrinum* produces the cholesterol lowering drug mevastatin, the world's first statin (Endo 2010). In contrast, *Aspergillus fumigatus* and *Aspergillus flavus* are pathogens, allergens, and mycotoxin producers[13,14] and *Penicillium expansum, Penicillium digitatum,* and *Penicillium italicum* are post-harvest pathogens of citrus fruits, stored grains, and other cereal crops[15–17].

Much of the rich diversity of ecologies and wide impact on human affairs that Aspergillaceae exhibit has been attributed to the remarkable chemical diversity of secondary metabolites, small molecules that function as toxins, signaling molecules, and pigments, that organisms in this family produce[18–20]. For example, diminished global production of secondary metabolites in *A. nidulans* caused by knocking-out the master regulator of secondary metabolism, *laeA*, resulted in increased predation by the collembolan fungivore, *Folsomia candida*, which suggests that these compounds play defensive roles[21]. Other studies investigating single secondary metabolites have

shown that these small molecules often have biological activities that are either harmful or beneficial to human welfare. For example, the *A. fumigatus*-produced secondary metabolite gliotoxin is a potent virulence factor in cases of systemic mycosis in vertebrates[22], and the *A. flavus*-produced secondary metabolite aflatoxin is among the most toxic and carcinogenic naturally occurring compounds[19,23]. In contrast, other secondary metabolites are mainstay antibiotics and pharmaceuticals; for example, the *Penicillium chrysogenum*-produced penicillin is among the world's most widely used antibiotics[24–26] and the *P. citrinum*-produced cholesterol lowering statins are consistently among the world's blockbuster drugs[12].

Understanding the evolution of the diverse ecological lifestyles exhibited by Aspergillaceae members as well as the family's remarkable chemodiversity requires a robust phylogenetic framework. To date, most molecular phylogenies of the family Aspergillaceae are derived from single or few genes and have yielded conflicting results. For example, there is little consensus on whether the genus *Aspergillus* is monophyletic or if it includes species from other genera such as *Penicillium*[27,28]. Furthermore, studies using genome-scale amounts of data, which could have the power to resolve evolutionary relationships and identify underlying causes of conflict[29,30], have so far tended to use a small subset of fungi from either *Aspergillus* or *Penicillium*[31–33]. Additionally, these genome-scale studies typically build one phylogeny and, based on the high clade support values (e.g., bootstrap values) obtained, infer or assume that the topology obtained is highly accurate[31–34].

In very recent years, several phylogenomic analyses have shown that incongruence, the presence of topological conflict between different data sets or analyses, is widespread[29,35–37], and that

99 certain branches of the tree of life can be very challenging to resolve, even with genome-scale
100 amounts of data[38–42]. For example, analyses of the currently available genome-scale amounts of
101 data have not resolved the placement of the budding yeast family Ascoideaceae in the fungal
102 subphylum Saccharomycotina (phylum: Ascomycota)[38,42,43]. Comparison of the topologies
103 inferred in previous phylogenomic studies in Aspergillaceae[31–34] suggests the presence of
104 incongruence (Figure S1). For example, some studies have reported section *Nidulantes* to be the
105 sister group to section *Nigri*[31], whereas other studies have placed it as the sister group to
106 *Ochraceorosei*[33] (Figure S1).
107
108 To systematically evaluate the evolutionary relationships among Aspergillaceae and identify
109 instances of incongruence, we used the genome sequences of 81 fungi from Aspergillaceae
110 spanning 4 genera, 24 sections within *Aspergillus* and *Penicillium*, and 12 outgroup fungi to
111 construct nucleotide (NT) and amino acid (AA) versions of a 1,668-gene data matrix. Using
112 three different maximum likelihood schemes (i.e., gene-partitioned, unpartitioned, and
113 coalescence), we inferred phylogenies from the 1,668-gene data matrix as well as from five
114 additional 834-gene data matrices derived from the top 50% of genes harboring strong
115 phylogenetic signal according to five different criteria (alignment length, average bootstrap
116 value, taxon completeness, treeness / relative composition variability, and number of variable
117 sites). Comparisons of these phylogenies coupled with complementary measures of internode
118 certainty[29,44,45] identified 12 / 78 (15.4%) incongruent bipartitions in the phylogeny of
119 Aspergillaceae. These cases of incongruence can be grouped into three categories: (i) 2 shallow
120 bipartitions with low levels of incongruence likely driven by incomplete lineage sorting, (ii) 3
121 shallow bipartitions with high levels of incongruence likely driven by hybridization or

122 introgression (or very high levels of incomplete lineage sorting), and (iii) 7 deeper bipartitions
123 with varying levels of incongruence likely driven by reconstruction artifacts likely linked with
124 poor taxon sampling. We also estimated divergence times across Aspergillaceae using relaxed
125 molecular clock analyses. Our results suggest Aspergillaceae originated in the lower Cretaceous,
126 125.1 (95% Confidence Interval (CI): 146.7 - 102.1) million years ago (mya), and that
127 *Aspergillus* and *Penicillium* originated 84.3 mya (95% CI: 90.9 - 77.6) and 77.4 mya (95% CI:
128 94.0 - 61.0), respectively. We believe this phylogeny and timetree provides a state-of-the-art
129 platform for comparative genomic, ecological, and chemodiversity studies in this ecologically
130 diverse and biotechnologically and medically significant family of filamentous fungi.

Methods

Genome sequencing and assembly

Mycelia were grown on potato dextrose agar for 72 hours before lyophilization. Lyophilized mycelia were lysed by grinding in liquid nitrogen and suspension in extraction buffer (100 mM Tris-HCl pH 8, 250 mM NaCl, 50 mM EDTA, and 1% SDS). Genomic DNA was isolated from the lysate with a phenol/chloroform extraction followed by an ethanol precipitation.

DNA was sequenced with both paired-end and mate-pair strategies to generate a high-quality genome assembly. Paired-end libraries and Mate-pair libraries were constructed at the Genomics Services Lab at HudsonAlpha (Huntsville, Alabama) and sequenced on an Illumina HiSeq X sequencer. Paired-end libraries were constructed with the Illumina TruSeq DNA kit, and mate-pair libraries were constructed with the Illumina Nextera Mate Pair Library kit targeting an insert size of 4 Kb. In total, 63 million paired-end reads and 105 million mate-pair reads were generated.

The *A. delacroxii* genome was assembled using the iWGS pipeline[46]. Paired-end and mate-pair reads were assembled with SPAdes, version 3.6.2[47], using optimal k-mer lengths chosen using KmerGenie, version 1.6982[48] and evaluated with QUAST, version 3.2[49]. The resulting assembly is 33.8 MB in size with an N50 of 939 Kb.

Data collection and quality assessment

To collect a comprehensive set of genomes representative of Aspergillaceae, we used 'Aspergillaceae' as a search term in NCBI's Taxonomy Browser and downloaded a representative genome from every species that had a sequenced genome as of February 5th 2018. We next confirmed that each species belonged to Aspergillaceae according to previous literature reports[31,50]. Altogether, 80 publicly available genomes and 1 newly sequenced genome spanning 5 genera (45 *Aspergillus* species; 33 *Penicillium* species; one *Xeromyces* species; one *Monascus* species; and one *Penicilliopsis* species) from the family Aspergillaceae were collected (File S1). We also retrieved an additional 12 fungal genomes from representative species in the order Eurotiales but outside the family Aspergillaceae to use as outgroups.

To determine if the genomes contained gene sets of sufficient quality for use in phylogenomic analyses, we examined their gene set completeness using Benchmarking Universal Single-Copy Orthologs (BUSCO), version 2.0.1[51] (Figure S2). In brief, BUSCO uses a consensus sequence built from hidden Markov models derived from 50 different fungal species using HMMER, version 3.1b2[52] as a query in TBLASTN[53,54] to search an individual genome for 3,156 predefined orthologs (referred to as BUSCO genes) from the Pezizomycotina database (creation date: 02-13-2016) available from ORTHODB, version 9[55]. To determine the copy number and completeness of each BUSCO gene in a genome, gene structure is predicted using AUGUSTUS, version 2.5.5[56], from the nucleotide coordinates of putative genes identified using BLAST and then aligned to the HMM alignment of the same BUSCO gene. Genes are considered "single copy" if there is only one complete predicted gene present in the genome, "duplicated" if there are two or more complete predicted genes for one BUSCO gene, "fragmented" if the predicted gene is

175 shorter than 95% of the aligned sequence lengths from the 50 different fungal species, and
176 "missing" if there is no predicted gene.
177

178 **Phylogenomic data matrix construction**

179 In addition to their utility as a measure of genome completeness, BUSCO genes have also proven
180 to be useful markers for phylogenomic inference[57], and have been successfully used in
181 phylogenomic studies of clades spanning the tree of life, such as birds[58], insects[59], and budding
182 yeasts[38]. To infer evolutionary relationships, we constructed nucleotide (NT) and amino acid
183 (AA) versions of a data matrix comprised of the aligned and trimmed sequences of numerous
184 BUSCO genes (Figure S3). To construct this data matrix, we first used the BUSCO output
185 summary files to identify orthologous single copy BUSCO genes with > 50% taxon-occupancy
186 (i.e., greater than 47 / 93 taxa have the BUSCO gene present in their genome); 3,138 (99.4%)
187 BUSCO genes met this criterion. For each BUSCO gene, we next created individual AA fasta
188 files by combining sequences across all taxa that have the BUSCO gene present. For each gene
189 individually, we aligned the sequences in the AA fasta file using MAFFT, version 7.294b[60], with
190 the BLOSUM62 matrix of substitutions[61], a gap penalty of 1.0, 1,000 maximum iterations, and
191 the 'genafpair' parameter. To create a codon-based alignment, we used a custom PYTHON,
192 version 3.5.2 (https://www.python.org/), script using BIOPYTHON, version 1.7[62], to thread codons
193 onto the AA alignment. The NT and AA sequences were then individually trimmed using
194 TRIMAL, version 1.4 [63], with the 'automated1' parameter. We next removed BUSCO genes
195 whose sequence lengths were less than 50% of the untrimmed length in either the NT or AA
196 sequences resulting in 1,773 (56.2%) BUSCO genes. Lastly, we removed BUSCO genes whose
197 trimmed sequence lengths were too short (defined as genes whose alignment length was less than

198 or equal to 167 AAs and 501 NTs), resulting in 1,668 (52.9%) BUSCO genes. The NT and AA
199 alignments of these 1,668 BUSCO genes were then concatenated into the full 1,668-gene NT and
200 AA versions of the phylogenomic data matrix.
201
202 To examine the stability of inferred relationships across all taxa, we constructed additional NT
203 and AA data matrices by subsampling genes from the 1,668-gene data matrix that harbor
204 signatures of strong phylogenetic signal. More specifically, we used 5 measures associated with
205 strong phylogenetic signal[64] to create 5 additional data matrices (1 data matrix per measure)
206 comprised of the top scoring 834 (50%) genes for NTs and AAs (Figure S4). These five
207 measures were: alignment length, average bootstrap value, taxon completeness, treeness /
208 relative composition variability (RCV)[65], and the number of variable sites. We calculated each
209 measure with custom PYTHON scripts using BIOPYTHON. Treeness / RCV was calculated using
210 the following formula:

211 $$\frac{Treeness}{RCV} = \frac{\sum_{u=1}^{b} l_u / l_t}{\sum_{i=1}^{c}\sum_{j=1}^{n} \frac{|c_{ij} - \overline{c_i}|}{s \bullet n}}$$

212 where l_u refers to the internal branch length of the *u*th branch (of *b* internal branches), l_t refers to
213 total tree length, *c* is the number of different characters per sequence type (4 for nucleotides and
214 20 for amino acids), *n* is the number of taxa in the alignment, c_{ij} refers to the number of *i*th *c*
215 characters for the *j*th taxon, $\overline{c_i}$ refers to the average number of the *i*th *c* character across *n* taxa,
216 and *s* refers to the total number of sites in the alignment. Altogether, we constructed a total of 12
217 data matrices (one 1,668-gene NT data matrix, one 1,668-gene AA data matrix, five NT
218 subsample data matrices, and five AA subsample data matrices).
219

Maximum likelihood phylogenetic analyses

We implemented a maximum likelihood framework to infer evolutionary relationships among taxa for each of the 1,668 single genes and each of the 12 data matrices separately. For inferences made using either the 1,668- or 834-gene data matrices, we used three different analytical schemes: concatenation with gene-based partitioning, concatenation without partitioning, and gene-based coalescence[30,66–68]. All phylogenetic trees were built using IQ-TREE, version 1.6.1[69]. In each case, we first determined the best model for each single gene or partition using the "-m TEST" parameter, which automatically estimates the best fitting model of substitutions according to their Bayesian Information Criterion values for either NTs or AAs[70]. Because we were unsure if downstream analyses may include the use of RAxML[71], we restricted the models tested and used to those shared by RAxML[71] and IQ-TREE by using the "-mset" parameter.

We first examined the inferred best fitting models across all single gene trees. Among NT genes, the best fitting model for 1,643 genes was a general time reversible model with unequal rates and unequal base frequencies with discrete gamma models, "GTR+G4"[72–74], and for the remaining 25 genes was a general time reversible model with invariable sites plus discrete gamma models, "GTR+I+G4"[74,75] (Figure S5a). Among AA genes, the best fitting model for 643 genes was the JTT model with invariable sites plus discrete gamma models, "JTT+I+G4"[75,76], for 362 genes was the LG model with invariable sites and discrete gamma models, "LG+I+G4"[75,77], for 225 genes was the JTT model with invariable sites, empirical AA frequencies, and discrete gamma models "JTT+F+I+G4"[75,76], and for 153 genes was the JTTDCMut model with invariable sites and discrete gamma models, "JTTDCMut+I+G4"[75,78] (Figure S5b).

243

244 We used IQ-TREE for downstream analysis because a recent study using diverse empirical
245 phylogenomic data matrices showed that it is a top-performing software[79] as well as because IQ-
246 TREE's gene partitioning scheme can account for different models of rate heterogeneity per
247 gene[80].

248

249 To determine the phylogeny of Aspergillaceae using a partitioned scheme where each gene has
250 its own model of sequence substitution and rate heterogeneity parameters, we created an
251 additional input file describing these and gene boundary parameters. More specifically, we
252 created a nexus-style partition file that was used as input with the "-spp" parameter[80]. To
253 increase the number of candidate trees used during maximum likelihood search, we set the "-
254 nbest" parameter to 10. Lastly, we conducted 5 independent searches for the maximum
255 likelihood topology using 5 distinct seeds specified with the "-seed" parameter and chose the
256 search with the best log-likelihood score. We used the phylogeny inferred using a partitioned
257 scheme on the full NT data matrix as the reference one for all subsequent comparisons (Figure
258 1).

259

260 To determine the phylogeny of Aspergillaceae using a non-partitioned scheme, we used all the
261 same parameters as above; the only difference was that we used a single model of sequence
262 substitution and rate heterogeneity parameters across the entire matrix. The most appropriate
263 single model was determined by counting which best fitting model was most commonly
264 observed across single gene trees. The most commonly observed model was
265 "GTR+F+I+G4"[75,81], which was favored in 1,643 / 1,668 (98.5%) of single genes, and

266 "JTT+I+G4"[75,76], which was favored in 643 / 1,668 (38.5%) of single genes, for NTs and AAs,
267 respectively, (Figure S5). In each analysis, the chosen model was specified using the "-m"
268 parameter.
269
270 To determine the phylogeny of Aspergillaceae using coalescence, a method that estimates
271 species phylogeny from single gene trees under the multi-species coalescent[67], we combined all
272 NEWICK[82,83] formatted single gene trees inferred using their best fitting models into a single file.
273 The resulting file was used as input to ASTRAL-II, version 4.10.12[68] with default parameters.
274
275 To evaluate support for single gene trees and for the reference phylogeny (Figure 1), we used an
276 ultrafast bootstrap approximation approach (UFBoot)[84]. UFBoot first generates bootstrap
277 alignments and creates an initial set of trees to use as a null distribution of starting trees. UFBoot
278 then uses quartet puzzling and the NNI algorithm[85,86] to sample the local maxima and their
279 neighborhoods in tree space while reducing run-time by re-estimating log-likelihood threshold
280 values to ensure only trees with sufficiently high log-likelihood values are investigated. If a new
281 tree exceeds the log-likelihood minimum, which is adaptively estimated based on the number of
282 trees encountered and the number of iterations performed by the quartet puzzling and NNI
283 algorithm, a resampling estimated log-likelihood score[87,88] is determined for the new tree. If the
284 resampling estimated log-likelihood score is better than the previous tree, the previous tree is
285 replaced with the new tree for the particular bootstrap alignment. Ultimately, this methodology is
286 3.1-10.2 times faster than rapid bootstrap support[89], is robust to moderate model violations, and,
287 most importantly, generates results that are unbiased compared to classic bootstrapping
288 techniques[84,90]. Thus, this method allows for a fast and accurate alternative to the classic

bootstrapping approach. To implement UFBoot for the NT 1,668-gene data matrix and single gene trees, we used the "-bb" option in IQ-TREE with 5,000 and 2,000 ultrafast bootstrap replicates, respectively.

Evaluating topological support

To identify and quantify incongruence, we used two approaches. In the first approach, we compared the 36 topologies inferred from the full 1,668-gene NT and AA data matrices and five additional 834-gene data matrices (constructed by selecting the genes that have the highest scores in five measures previously shown to be associated with strong phylogenetic signal; see above) using three different maximum likelihood schemes (i.e., gene partitioned, non-partitioned, coalescence) and identified all incongruent bipartitions between the reference phylogeny (Figure 1) and the other 35. In the second approach, we scrutinized each bipartition in the reference phylogeny using measures of internode certainty (IC) measures for complete and partial single gene trees[29,44,45]. To better understand single gene support among conflicting bipartitions, we calculated gene-wise log-likelihood scores (GLS)[42] and gene support frequencies (GSF) for the reference and alternative topologies at conflicting bipartitions.

Identifying internodes with conflict across subsampled data matrices

To identify incongruent bipartitions between the reference phylogeny and the other 35 phylogenies, we first included the 36 generated phylogenetic trees into a single file. We next evaluated the support of all bipartitions in the reference topology among the other 35 phylogenies using the "-z" option in RAxML. Any bipartition in the reference phylogeny that was not present in the rest was considered incongruent; each conflicting bipartition was

identified through manual examination of the conflicting phylogenies. To determine if sequence type, subsampling method, or maximum likelihood scheme was contributing to differences in observed topologies among conflicting internodes, we conducted multiple correspondence analysis of these features among the 36 phylogenies and visualized results using the R, version 3.3.2[91], packages FACTOMINER, version 1.40[92] and FACTOEXTRA, version 1.0.5[93].

Identifying internodes with conflict across the 1,668 gene trees

To examine the presence and degree of support for bipartitions that conflict with the bipartitions in a given phylogeny, we calculated the internode certainty[29,44,45,94] of all internodes in the reference phylogeny (Figure 1) using the 1,668 gene trees as input. In general, IC scores near 0 indicate that there is near-equal support for an alternative, conflicting bipartition among a set of trees compared to a given bipartition present in the reference topology, which is indicative of high conflict. Therefore, we investigated incongruence in all internodes in the reference phylogeny (Figure 1) that exhibited IC scores lower than 0.1. To calculate IC values for each bipartition for the reference phylogeny, we created a file with all 1,668 complete and partial single gene trees. The resulting file of gene trees, specified with the "-z" parameter in RAXML, were used to calculate IC values using the "-f i" argument. The topology was specified with the "-t" parameter. Lastly, we used the Lossless corrected IC scoring scheme, which corrects for variation in taxon number across single gene trees[44]. We also used these IC values to inform which data type (NT or AA) provided the strongest signal for the given set of taxa and sequences. We observed that NTs consistently exhibited higher IC scores than AAs (hence our decision to use the topology inferred from the full NT data matrix using a gene-partitioned scheme – shown in Figure 1 – as the 'reference' topology in all downstream analyses).

335

336 *Examining gene-wise log-likelihood scores for incongruent internodes*

337 To determine the per gene distribution of phylogenetic signal supporting a bipartition in the
338 reference phylogeny or a conflicting bipartition, we calculated gene-wise log-likelihood scores
339 (GLS)[42] using the NT data matrix. We chose to calculate GLS using the NT data matrix because
340 distributions of IC values from phylogenies inferred using NTs had consistently higher IC values
341 across schemes and data matrices (Figure S6). To do so, we used functions available in IQ-
342 TREE. More specifically, we inputted a phylogeny with the reference or alternative topology
343 using the "-te" parameter and informed IQ-TREE of gene boundaries, their corresponding
344 models, and optimal rate heterogeneity parameters in the full 1,668-gene data matrix using the "-
345 spp" parameter. Lastly, we specified that partition log-likelihoods be outputted using the "-wpl"
346 parameter. To determine if a gene provided greater support for the reference or alternative
347 bipartition, we calculated the difference in GLS (ΔGLS) using the following formula:

348 $$\Delta\text{GLS}_i = \ln L(G_i)_{ref} - \ln L(G_i)_{alt}$$

349 where $ln\ L(G_i)_{ref}$ and $ln\ L(G_i)_{alt}$ represent the log-likelihood values for the reference and
350 alternative topologies for gene G_i. Thus, values greater than 0 reflect genes in favor of the
351 reference bipartition, values lower than 0 reflect genes in favor of the alternative bipartition, and
352 values of 0 reflect equal support between the reference and alternative bipartitions.

353

354 *Calculating gene support frequencies for reference and conflicting bipartitions*

355 We next examined support for bipartitions in the reference topology as well as for their most
356 prevalent conflicting bipartitions by calculating their gene support frequencies (GSF). GSF refers
357 to the fraction of single gene trees that recover a particular bipartition. Currently, RAxML can

only calculate GSF for trees with full taxon representation. Since our dataset contained partial gene trees, we conducted customs tests for determining GSF. To calculate GSF for NT (GSF_{NT}) and AA (GSF_{AA}) single gene trees, we extracted subtrees for the taxa of interest in individual single gene trees and counted the occurrence of various topologies. For example, consider there are three taxa represented as A, B, and C, the reference rooted topology is "((A,B),C);" and the alternative rooted topology is "((A,C),B);". We counted how many single gene trees supported "(A,B)," or "(A, C),". For reference and alternative topologies involving more than three taxa or sections, we conducted similar tests. For example, if the reference rooted topology is "(((A,B),C),D);" and the alternative rooted topology is "((A,B),(C,D));", we counted how many single gene phylogenies supported "((A,B),C)," as sister to D and how many single gene phylogenies supported "(A,B)," and "(C,D)," as pairs of sister clades. For conflicting bipartitions at shallow depths in the phylogeny (i.e., among closely related species), we required all taxa to be present in a single gene tree; for conflicting bipartitions near the base of the phylogeny (i.e., typically involving multiple sections), we required at least one species to be present from each section of interest (with the exception of *Exilicaulis* because this section is not monophyletic). Scripts to determine GSF were written using functions provided in NEWICK UTILITIES, version 1.6[95].

Estimating divergence times

To estimate the divergence times for the phylogeny of the Aspergillaceae, we analyzed our NT data matrix used the Bayesian method implemented in MCMCTREE from the PAML package, version 4.9d[96]. To do so, we conducted four analyses: we (i) identified genes evolving in a "clock-like" manner from the full data matrix, (ii) estimated the substitution rate across these

genes, (iii) estimated the gradient and Hessian[97] at the maximum likelihood estimates of branch lengths, and (iv) estimated divergence times by Markov chain Monte Carlo (MCMC) analysis.

(i) Identifying "clock-like" genes

Currently, large phylogenomic data matrices that contain hundreds to thousands of genes and many dozens of taxa are intractable for Bayesian inference of divergence times; thus, we identified and used only those genes that appear to have evolved in a "clock-like" manner in the inference of divergence times. To identify genes evolving in a "clock-like" manner, we calculated the degree of violation of a molecular clock (DVMC)[98] for single gene trees. DVMC is the standard deviation of root to tip distances in a phylogeny and is calculated using the following formula:

$$\mathrm{DVMC} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(t_i - \bar{t})^2}$$

where t_i represents the distance between the root and species i across n species. Using this method, genes with low DVMC values evolve in a "clock-like" manner compared to those with higher values. We took the top scoring 834 (50%) genes and bootstrap subsampled 250 genes without replacement. We decided to use 250 genes because a previous study with a similar number of taxa used a similar number of genes[98].

(ii) Estimating substitution rate

To estimate the substitution rate across the 250 genes, we used BASEML from the PAML package, version 4.9d[96]. We estimated substitution rate using a "GTR+G" model of substitutions (model = 7) and a strict clock model (clock = 1). Additionally, we point calibrated the root of the tree to 96

403 million years ago (mya) according to TIMETREE[99], which is based on previous estimations [100]:
404 50.0 mya; [101]: 96.1 mya; [102]: 146.1 mya. We found the estimated per site substitution rate per
405 time unit was 0.04.
406
407 (iii) Estimation of the gradient and Hessian
408 To save computing time, the likelihood of the alignment was approximated using a gradient and
409 Hessian matrix. The gradient and Hessian refer to the first and second derivatives of the log-
410 likelihood function at the maximum likelihood estimates of branch lengths[97], and collectively
411 describe the curvature of the log-likelihood surface. Estimating gradient and Hessian requires an
412 input tree with specified time constraints. For time constraints, we used the *Aspergillus flavus –*
413 *Aspergillus oryzae* split (3.68-3.99 mya[102,103]), the *Aspergillus fumigatus – Aspergillus clavatus*
414 split (35-59 mya[102,103]), the origin of the genus *Aspergillus* (43-85 mya[102,104–107]), and the origin
415 of Aspergillaceae (50-146 mya[100–102]) as obtained from TIMETREE[99].
416
417 (iv) Estimating divergence times using MCMC analysis
418 To estimate divergence times, we used the resulting gradient and Hessian results from the
419 previous step for use in MCMC analysis using MCMCTREE[96]. To do so, a gamma distribution
420 prior shape and scale must be specified. The gamma distribution shape and scale is determined
421 from the substitution rate determined in step ii where shape is $a=(s/s)^2$ and scale is $b=s/s^2$ and s
422 is the substitution rate. Therefore, $a=1$ and $b=25$ and the "rgene_gamma" parameter was set to
423 "1 25." We also set the "sigma2_gamma" parameter to "1 4.5." To minimize the effect of initial
424 values on the posterior inference, we discarded the first 100,000 results. Thereafter, we sampled
425 every 500 iterations until 10,000 samples were gathered. Altogether, we ran 5.1 million iterations

(100,000 + 500 x 10,000), which is 510 times greater than the recommended minimum for MCMC analysis[108]. Lastly, we set the "finetune" parameter to 1.

To determine the stability of inferred divergence time estimates, we constructed two additional matrices of 250 genes, repeated the analyses in steps ii-iv, and compared results. The two additional data matrices were constructed by independently bootstrap subsampling 250 genes without replacement from the 834 genes with the best DVMC values. We subsequently repeated steps ii-iv and conducted correlation analyses between the three sets of 250 genes to determine the stability of inferred divergence times.

Statistical analysis and figure making

All statistical analyses were conducted in R, version 3.3.2[91]. Spearman rank correlation analyses[109] were conducted using the "rcorr" function in the package HMISC, version 4.1-1[110]. Stacked barplots, barplots, histograms, scatterplots, and boxplots were made using GGPLOT2, version 2.2.1[111]. Intersection plots (also known as UpSet plots), were made using UPSETR, version 1.3.3[112]. The topological similarity heatmap and hierarchical clustering was done using PHEATMAP, version 1.0.8[113]. Phylogenetic trees were visualized using FIGTREE, version 1.4.3[114]. The phylogenetic tree with the geological time scale was visualized using STRAP, version 1.4[115]. Artistic features of figures (e.g., font size, font style, etc.) were minimally edited using the graphic design software Affinity Designer (https://affinity.serif.com/en-us/).

Results

The examined genomes have nearly complete gene sets

Assessment of individual gene set completeness showed that most of the 93 genomes (81 in the ingroup and 12 in the outgroup) used in our study contain nearly complete gene sets and that all 93 genomes are appropriate for phylogenomic analyses. Specifically, the average percentage of BUSCO single-copy genes from the Pezizomycotina database[55] present was 96.2 ± 2.6% (minimum: 81.1%; maximum: 98.9%; Figure S2). Across the 93 genomes, only 3 (3.2%) genomes had < 90% of the BUSCO genes present in single-copy (*Penicillium carneum*: 88.6%; *Penicillium verrucosum*: 86.1%; and *Histoplasma capsulatum*: 81.1%).

The generated data matrices exhibit very high taxon occupancy

The NT and AA alignments of the 1,668-gene data matrix were comprised of 3,163,258 and 1,054,025 sites, respectively. The data matrix exhibited very high taxon occupancy (average gene taxon occupancy: 97.2 ± 0.1%; minimum: 52.7%; maximum: 100%; Figure S7a, b; File S2). 417 genes had 100% taxon-occupancy, 1,176 genes had taxon-occupancy in the 90% to 99.9% range, and only 75 genes had taxon occupancy lower than 90%. Assessment of the 1,668 genes for five criteria associated with strong phylogenetic signal (gene-wise alignment length, average bootstrap value, completeness, treeness / RCV, and the number of variable sites) facilitated the construction of five subsampled matrices derived from 50% of the top scoring genes (Figure S7; File S2).

Examination of the gene content differences between the 5 NT subsampled data matrices as well as between the 5 AA data matrices revealed that they are composed of variable sets of genes

(Figure S8). For example, the largest intersection among NT data matrices comprised of 207 genes that were shared between all NT matrices except the completeness-based one; similarly, the largest intersection among AA data matrices was 228 genes and was shared between all AA matrices except the completeness-based one (Figure S8a, b). Examination of the number of gene overlap between the NT and AA data matrices for each criterion (Figure S8c) showed that three criteria yielded identical or nearly identical NT and AA gene sets. These were completeness (834 / 834; 100% shared genes; $r_s = 1.00$, $p < 0.01$; Figure S7c), alignment length (829 / 834; 99.4% shared genes; $r_s = 1.00$, $p < 0.01$; Figure S7f), and the number of variable sites (798 / 834; 95.7% shared genes; $r_s = 0.99$, $p < 0.01$; Figure S7i). The other two criteria showed greater differences between NT and AA data matrices (average bootstrap value: 667 / 834; 80.0% shared genes; $r_s = 0.78$, $p < 0.01$; Figure S7l; treeness / RCV: 644 / 834; 77.2% shared genes; $r_s = 0.72$, $p < 0.01$; Figure S7o).

A genome-scale phylogeny for the family Aspergillaceae

NT and AA phylogenomic analyses of the full data matrix and the five subsampled data matrices under three analytical schemes recovered a broadly consistent set of relationships (Figure 1, 2, 3, 4). Across all 36 species-level phylogenies, we observed high levels of topological similarity (average topological similarity: 97.2 ± 2.5%; minimum: 92.2%; maximum: 100%) (Figure 2), with both major genera (*Aspergillus* and *Penicillium*) as well as all sections, with the exception of *Exilicaulis*, in *Aspergillus* and *Penicillium*[50,116] recovered as monophyletic (Figures 1, 3, and 4). Additionally, all but one internodes exhibited absolute UFBoot scores[84]; the sole exception was internode 33 (I33), which received 95 UFBoot support (Figure 1 and S9).

Surprisingly, one taxon previously reported to be part of Aspergillaceae, *Basipetospora chlamydospora*, was consistently placed among outgroup species (Figure 1) and may represent a misidentified isolate. A similarly surprising placement was observed for *Aspergillus ochraceoroseus* IBT 24754[33], which our phylogenies consistently placed in section *Nigri* (Figure 1) rather than, as expected based on previous work, in section *Ochraceorosei*[117]. To explore this placement further, we reconstructed a phylogeny of closely related *Aspergillus* species from sections *Flavi, Ochraceorosei, Usti, Versicolores, Nidulantes,* and *Nigri* and included another *A. ochraceoroseus* isolate, strain SRRC1432[118] using the same set of 1,668 BUSCO genes as well as a larger set of 3,150 BUSCO genes. Phylogenomic analysis of these two data matrices recovered *A. ochraceoroseus* SRRC1432 as sister to *A. rambellii* in section *Ochraceorosei*, consistent with the original description of section *Ochraceorosei*[119]. In contrast, *A. ochraceoroseus* IBT 24754 remained placed within *Nigri* (Figure S10a and b). Hypothesizing that *A. ochraceoroseus* IBT 24754 may represent a misidentified isolate, we examined its genome size and number of genes in relation to those of *A. ochraceoroseus* SRRC1432 and *A. rambellii* and found them to be very different (Figure S10c). Specifically, *A. ochraceoroseus* IBT 24754 has 11,939 genes and a genome size of 35.4 Mbp while *A. ochraceoroseus* SRRC1432 and *A. rambellii* have gene counts of 7,829 and 7,761 and genome sizes of 24.3 and 26.4 Mbp. Together, these results suggest that *A. ochraceoroseus* IBT 24754 is a misidentified *Aspergillus* species belonging to section *Nigri*; to avoid further confusion, we henceforth refer to this strain as *A. spp.* IBT 24574 (Figure S10d).

Examination of the Aspergillaceae phylogeny reveals 12 incongruent bipartitions

Examination of all 36 species-level phylogenies revealed the existence of 8 (8 / 78; 10.3%) incongruent bipartitions. Complementary examination of IC, a bipartition-based measure of incongruence, revealed an additional 4 / 78 (5.1%) bipartitions that displayed very high levels of incongruence at the gene level, raising the total number of incongruent bipartitions to 12 (12 / 78; 15.4%).

Examination of the eight conflicting bipartitions stemming from the comparison of the 36 phylogenies showed that they were very often associated with data type (NT or AA) and scheme employed (concatenation or coalescence). For example, the first instance of incongruence concerns the identity of the sister species to *Penicillium biforme* (I60; Figure 1 and 3a); this species is *P. camemberti* in the reference phylogeny but analyses of the full and two subsampled AA data matrices with coalescence recover instead *Penicillium fuscoglaucum*. The data type and analytical scheme employed also appear to underlie the second and third instances of incongruence, which concern the polyphyly of section *Exilicaulis* (I74 and I78; Figures 1 and 3b), the fourth and fifth instances, which concern relationships among *Aspergillus* sections (I24 and I35; Figures 1 and 3c), as well as the sixth instance, which concerns relationships among the sections *Digitata*, *Chrysogena*, and *Roquefortorum* (I63; Figure 1 and 3d). The seventh instance is also associated with data type, but not with the scheme employed; while the reference as well as most subsampled NT matrices support the *Aspergillus persii* and *Aspergillus sclerotiorum* clade as sister to *Aspergillus westerdijkiae* (I33; Figure 1 and 3e), most AA data matrices recover a conflicting bipartition where *A. steynii* is the sister group of *A. westerdijkiae*. The final instance of incongruence was the least well supported, as 35 / 36 (97.2%) phylogenies supported

536 *Aspergillus kawachii* as the sister group to *Aspergillus awamori* (I15, Figure 1 and 3f), but
537 analysis of one AA subsampled data matrix with coalescence instead recovered *Aspergillus*
538 *luchuensis* as the sister group.
539
540 For each of these bipartitions (Figure 3), we examined clustering patterns using multiple
541 correspondence analysis of matrix features (i.e., sequence type and subsampling method) and
542 analysis scheme among trees that support the reference and alternative topologies (Figure S11).
543 Distinct clustering patterns were observed for I74, I78, and I33 (Figure 3 and S11). For I74 and
544 I78, there are three alternative, conflicting topologies, with the first two clustering separately
545 from the third (Figure 3b and S11b). For I33, phylogenies that support the reference and
546 alternative topologies formed distinct clusters (Figure 3e). Examination of the contribution of
547 variables along the second dimension, which is the one that differentiated variables that
548 supported each topology, revealed that the distinct clustering patterns were driven by sequence
549 type (Figure S11g and h).
550
551 Examination of IC values revealed four additional bipartitions with strong signatures for
552 incongruence at the gene level, defined as IC score lower than 0.10. The first instance concerns
553 the sister taxon to the *Aspergillus* and *Penicillium* clade. Although all 36 phylogenies recover a
554 clade comprised of *Xeromyces bisporus* and *Monascus ruber* as the sister group, the IC score for
555 this bipartition is 0.00 (I3; Figure 4a); the most prevalent, conflicting bipartition supports
556 *Penicilliopsis zonata* as sister to *Aspergillus* and *Penicillium* (Figure 4a). Similarly, although all
557 36 phylogenies recover *Penicillium* as sister to *Aspergillus*, the IC score for this bipartition is
558 also 0.00 (I4; Figure 4b); the most prevalent, conflicting bipartition supports *X. bisporus* and *M.*

559 *ruber* as the sister clade to *Aspergillus* (Figure 4b). In the third instance, all 36 phylogenies
560 support *Aspergillus novofumigatus* and *Aspergillus lentulus* as sister species, but the IC score of
561 this bipartition is 0.01 (I43; Figure 4c); the most prevalent, conflicting bipartition recovers *A.*
562 *lentulus* as the sister species to a clade comprised of *Aspergillus fumigatus* and *Aspergillus*
563 *fischeri* (Figure 4c). Finally, all 36 phylogenies supported a clade of *Penicillium solitum,*
564 *Penicillium polonicum*, and *Penicillium freii* as sister to a clade of *Penicillium nordicum* and
565 *Penicillium verrucosum*, but the IC score for this bipartition is 0.01 (I55; Figure 4d); the most
566 prevalent, conflicting bipartition supports the clade of *P. solitum, P. polonicum,* and *P. freii* as
567 sister to a clade of *P. camemberti, P. biforme* and *P. fuscoglaucum* (Figure 4d).
568
569 To examine the underlying individual gene support to the resolution of these 12 bipartitions, we
570 examined the phylogenetic signal contributed by each individual gene in the full NT data matrix.
571 In all 12 bipartitions, we found that inferences were robust to single gene outliers with strong
572 phylogenetic signal (Figure S12; File S4).
573
574 **Incongruence in the Aspergillaceae phylogeny**
575 Examination of the 12 incongruent bipartitions with respect to their placement on the phylogeny
576 (shallow, i.e., near the tips of the phylogeny or deeper, i.e., away from the tips and toward the
577 base of the phylogeny) and the amount of conflict (quantified using IC and GSF) allowed us to
578 group them into three categories: (i) shallow bipartitions (I15 and I60) with low levels of
579 incongruence, (ii) shallow bipartitions (I33, I43, and I55) with high levels of incongruence, and
580 (iii) deeper bipartitions (I3, I4, I24, I35, I63, I74, and I78) with varying levels of incongruence
581 and typically associated with single taxon long branches.

582

583 (i) Shallow bipartitions with low levels of incongruence

584 The two bipartitions that fell into this category, I60 (Figure 3a) and I15 (Figure 3f), exhibited
585 low levels of incongruence among closely related taxa. For I60, the reference bipartition was
586 observed in 33 / 36 phylogenies, had an IC score of 0.22, and GSF_{NT} and GSF_{AA} scores of 0.70
587 and 0.21, respectively. Similarly, the reference bipartition for I15 was observed in 35 / 36
588 phylogenies, had an IC score of 0.39, and GSF_{NT} and GSF_{AA} scores of 0.84 and 0.47,
589 respectively. Notably, the GSF_{NT} scores were substantially higher for the reference bipartitions in
590 both of these cases.

591

592 (ii) Shallow bipartitions with high levels of incongruence

593 The three shallow bipartitions, I33 (Figure 3e), I43 (Figure 4c), and I55 (Figure 4d), in this
594 category exhibited high levels of incongruence among closely related taxa. For I33, the reference
595 bipartition was observed in 16 / 36 (44.4%), had an IC score of 0.00, and GSF_{NT} and GSF_{AA}
596 scores of 0.38 and 0.27, respectively. The reference bipartition for I43 was observed in all 36
597 phylogenies, had an IC score of 0.01 and GSF_{NT} and GSF_{AA} scores of 0.39 and 0.22,
598 respectively. Similarly, the reference bipartition I55 was observed in all 36 phylogenies, had an
599 IC score of 0.01, and GSF_{NT} and GSF_{AA} scores of 0.51 and 0.31, respectively. Notably, in all
600 three cases, substantial fractions of genes supported both the reference and the conflicting
601 bipartitions, with both the GSF_{NT} and GSF_{AA} scores of each pair of bipartitions being almost
602 always higher than 0.2.

603

604 (iii) Deeper bipartitions often associated with single taxon long branches

605 The seven bipartitions in this category were I74 and I78 (Figure 3b), I24 and I35 (Figure 3c), I63
606 (Figure 3d), I3 (Figure 4a), and I4 (Figure 4b). All of them are located deeper in the tree and
607 most involve single taxa with long terminal branches (Figure 1). The reference bipartitions for
608 internodes I74 and I78, which concern relationships among the sections *Lanata-divaricata*,
609 *Exilicaulis* and *Citrina*, were observed in 26 / 36 (72.2%) phylogenies; the remaining 10 / 36
610 (27.8%) phylogenies recovered three alternative, conflicting bipartitions. Both reference
611 bipartitions had IC scores of 0.01, and GSF_{NT} and GSF_{AA} scores of 0.11 and 0.07, respectively.
612 The reference bipartitions for internodes I24 and I35, which concern the placement of
613 *Aspergillus terreus*, the single taxon representative of section *Terrei*, were observed in 27 / 36
614 (75.0%) phylogenies, had IC scores of 0.01 and 0.02, and GSF_{NT} and GSF_{AA} scores of 0.17 and
615 0.09, respectively. The reference bipartition I63, which involved the placement of the
616 *Penicillium digitatum*, the sole representative of section *Digitata*, was observed in 28 / 36
617 (77.8%), had an IC score of 0.07, and GSF_{NT} and GSF_{AA} scores of 0.41 and 0.28, respectively.
618 Finally, the reference bipartitions I3 and I4 (Figure 4), which concern the identity of the sister
619 taxon of *Aspergillus* and *Penicillium* (I3) and the identity of the sister taxon of *Aspergillus* (I4),
620 were not observed among the 36 phylogenies but both had IC values of 0.00. For I3, GSF_{NT} and
621 GSF_{AA} scores were 0.12 and 0.15, respectively. For I4, GSF_{NT} and GSF_{AA} scores were 0.24 and
622 0.28, respectively.
623
624 **A geological timeline for the evolutionary diversification of the Aspergillaceae**
625 **family**
626 To estimate the evolutionary diversification among *Aspergillaceae*, we subsampled the 1,668-
627 gene matrix for high-quality genes with "clock-like" rates of evolution by examining DVMC [98]

values among single gene trees. Examination of the DVMC values facilitated the identification of a tractable set of high-quality genes for relaxed molecular clock analyses (Figure S13). We found that *Aspergillaceae* originated 125.1 (95% CI: 146.7 - 102.1) mya during the Cretaceous period (Figure 5). We found that the common ancestor of *Aspergillus* and *Penicillium* split from the *X. bisporus* and *M. ruber* clade shortly thereafter, approximately 114.3 (95% CI: 135.5 - 96.5) mya. We also found that the genera *Aspergillus* and *Penicillium* split 102.4 (95% CI: 122.3 - 88.2) mya, with the last common ancestor of *Aspergillus* originating approximately 84.3 mya (95% CI: 90.9 - 77.6) and the last common ancestor of *Penicillium* originating approximately 77.4 mya (95% CI: 94.0 - 61.0).

Our analysis also provides estimates of the origin of various iconic sections within *Aspergillus* and *Penicillium*. Among *Aspergillus* sections, section *Nigri*, which includes the industrial workhorse *A. niger*, originated 51.6 (95% CI: 63.4 - 38.1) mya. Section *Flavi*, which includes the food fermenters *A. oryzae* and *A. sojae* and the plant pathogen *A. flavus*, originated 32.6 (95% CI: 45.5 - 22.4) mya. Additionally, section *Fumigati*, which includes the opportunistic human pathogen *A. fumigatus*, originated 17.4 (95% CI: 24.7 - 11.9) mya. Among *Penicillium* sections, section *Fasiculata*, which contains Camembert and Brie cheese producer *P. camemberti* and the ochratoxin A producer, *P. verrucosum*, originated 7.1 (95% CI: 10.9 - 4.1) mya. Section *Chrysogena,* which includes the antibiotic penicillin producing species *P. chrysogenum*, originated 6.4 (95% CI: 11.5 - 3.2) mya. Additionally, section *Citrina*, which contains *P. citrinum*, which the first statin was isolated from and is commonly associated with moldy citrus fruits (Endo et al. 1976), originated 32.4 (95% CI: 46.1 - 20.8) mya.

Discussion

Our analyses provide a robust evaluation of the evolutionary relationships and diversification among Aspergillaceae, a family of biotechnologically and medically significant fungi. We scrutinized our proposed reference phylogeny (Figure 1) against 35 other phylogenies recovered using all possible combinations of six multi-gene data matrices (full or subsamples thereof), three maximum likelihood schemes, and two sequence types and complemented this analysis with bi-partitioned based measures of support (Figures 1 and 2). Through these analyses, we found that 12 / 78 (15.4%) bipartitions were incongruent (Figure 3 and 4) and explored the characteristics as well as sources of these instances of incongruence. Finally, we placed the evolution and diversification of Aspergillaceae in the context of geological time.

Comparison of our 81-taxon, 1,668-gene phylogeny to a previous one based on a maximum likelihood analysis of 9 loci for 204 Aspergillaceae species[116], suggests that our analyses identified and strongly supported several new relationships and resolved previously low supported bipartitions (Figure 1, Figure S14). The robust resolution of our phylogeny is likely due to the very large size of our data matrix, both in terms of genes as well as in terms of sequence. For example, the placement of *Aspergillus* section *Nigri* has been unstable in previous phylogenomic analyses (Figure S1)[31,33,34], but our denser sampling of taxa in this section as well as inclusion of representative taxa from sections *Nidulantes, Versicolores, Usti*, and *Ochraceorosei* now provides strong support for the sister relationship of the *Aspergillus* section *Nigri* to sections *Nidulantes, Versicolores, Usti*, and *Ochraceorosei* (Figure 1).

673 However, our analysis also identified several relationships that exhibit high levels of
674 incongruence (Figures 3 and 4). In general, gene tree incongruence can stem from biological or
675 analytical factors[30,42]. Biological processes such as incomplete lineage-sorting (ILS)[121],
676 hybridization[122], gene duplication and subsequent loss[123], horizontal gene transfer[124], and natural
677 selection[125,126], can cause the histories of genes to differ from one another and from the species
678 phylogeny. Importantly, although the expected patterns of incongruence will be different for
679 each factor and depend on a number of parameters, the observed patterns of conflict in each of
680 the 12 cases of incongruence in the Aspergillaceae phylogeny can yield insights and allow the
681 formation of hypotheses about the potential drivers in each case. For example, ILS often results
682 in relatively low levels of incongruence; for instance, examination of the human, chimp, and
683 gorilla genomes has showed that 20-25% of the gene histories differ from the species
684 phylogeny[127,128]. In contrast, recent hybridization is expected to typically produce much higher
685 levels of incongruence due to rampant sequence similarity among large amounts of genomic
686 content; for instance, examination of *Heliconius* butterfly genomes revealed incongruence levels
687 higher than 40%[129].
688
689 Additionally, analytical factors such as model choice[130] and taxon sampling[131,132] can lead to
690 erroneous inference of gene histories. Perhaps the most well-known instance of incongruence
691 stemming from analytical factors is what is known as “long branch attraction”, namely the
692 situation where highly divergent taxa, i.e., the ones with the longest branches in the phylogeny,
693 will often artifactually group with other long branches[133].
694

Examination of the patterns of incongruence in the Aspergillaceae phylogeny allows us to not only group the 12 incongruent internodes with respect to their patterns of conflict but also to postulate putative drivers of the observed incongruence. For example, both I15 and I60 are shallow internodes exhibiting low levels of incongruence, suggesting that one likely driver of the observed incongruence is ILS. In contrast, the shallow internodes I33, I43, and I55 exhibit much higher levels of incongruence that are most likely to be the end result of processes, such as hybridization or repeated introgression. Finally, the remaining seven incongruent internodes (I3, I4, I24, I35, I63, I74, and I78) exhibit varying levels of incongruence and are typically associated with single taxon long branches (Figures 1, 3, and 4), implicating taxon sampling as a likely driver of the observed incongruence. Given that inclusion of additional taxa robustly resolved the previously ambiguous placement of the long-branched *Aspergillus* section *Nigri* (see discussion above), we predict that additional sampling of taxa that break up the long branches associated with these seven internodes will lead to their robust resolution.

Finally, our relaxed molecular clock analysis of the Aspergillaceae phylogeny provides a robust but also comprehensive time-scale for the evolution of Aspergillaceae and its two large genera, *Aspergillus* and *Penicillium* (Figure 5), filling a gap in the literature. Previous molecular clock studies provided estimates for only four internodes, mostly within the genus *Aspergillus*[99–107] and yielded much greater time intervals. For example, the previous estimate for the origin of Aspergillaceae spanned nearly 100 mya (50-146 mya[100–102]) while our dataset and analysis provided a much narrower range of 44.5 mya (mean: 125.1; 95% CI: 146.7 - 102.1).

Conclusion

Fungi from Aspergillaceae have diverse ecologies and play significant roles in biotechnology and medicine. Although most of the 81 genomes from Aspergillaceae are skewed towards two iconic genera, *Aspergillus* and *Penicillium*, and do not fully reflect the diversity of the family, they do provide a unique opportunity to examine the evolutionary history of these important fungi using a phylogenomic approach. Our scrutiny of the Aspergillaceae phylogeny, from the Cretaceous to the present, provides strong support for most relationships within the family as well as identifies a few that deserve further examination. Our results suggest that the observed incongruence is likely associated with diverse processes such as incomplete lineage sorting, hybridization and introgression, as well as with analytical issues associated with poor taxon sampling. Our elucidation of the tempo and pattern of the evolutionary history of Aspergillaceae provides a robust phylogenetic and temporal framework for investigation the evolution of pathogenesis, secondary metabolism, and ecology of this diverse and important fungal family.

Data availability

All data matrices, species-level and single-gene phylogenies will be available through the figshare repository upon acceptance for publication. The genome sequence and raw reads of *Aspergillus delacroxii* have been uploaded to GenBank as BioProject PRJNA481010.

734 # Acknowledgements

735 We thank members of the Rokas laboratory for helpful suggestions and discussion. We thank
736 Nathan McDonald for technical help with DNA extraction. This work was conducted, in part,
737 using the Advanced Computing Center for Research and Education at Vanderbilt University. JLS
738 was supported by the Graduate Program in Biological Sciences at Vanderbilt University. This
739 work was supported, in part, by the National Science Foundation (DEB-1442113 to A.R.), the
740 Burroughs Wellcome Fund, and a Guggenheim fellowship (to A.R.).

741

Main Figure Legends

Figure 1. A robust genome-scale phylogeny for the fungal family Aspergillaceae. Different genera are depicted using different colored boxes; *Aspergillus* is shown in red, *Penicillium* in blue, *Xeromyces* in green, *Monascus* in purple, and *Penicilliopsis* in orange. Different sections within *Aspergillus* and *Penicillium* are depicted with alternating dark grey and grey bars. Internode certainty values are shown below each internode and bootstrap values are shown above each internode (only bootstrap values lower than 100 percent supported are shown). Internode certainty values were calculated using the 1,668 maximum likelihood single gene trees. 5,000 ultrafast bootstrap replicates were used to determine internode support. Internodes were considered unresolved if they were not present in one or more of the other 35 phylogenies represented in Figure 2 – the branches of these unresolved internodes are drawn in red. The inset depicts the phylogeny with branch lengths corresponding to estimated nucleotide substitutions per site. Colored circles next to species names indicate the lifestyle or utility of the species (i.e., animal pathogen, dark orange; plant pathogen, purple; food fermenter, green; post-harvest food contaminant, pink; industrial workhorse, grey; genetic model, black; other, white). Exemplary secondary metabolites produced by different Aspergillaceae species are depicted to the right of the colored circles.

Figure 2. Topological similarity between the 36 phylogenies constructed using 6 different data matrices, 2 different sequence types, and 3 analytical schemes. (a) A heatmap depiction of topological similarity between the 36 phylogenies constructed in this study. The 36 phylogenies were inferred from analyses of 2 different sequence types (i.e., protein: depicted in black; nucleotide: depicted in white), 3 different analytical schemes (i.e.,

765 partitioned: depicted in black; non-partitioned: depicted in grey; coalescence: depicted in white),
766 and 6 different matrices (full data matrix: "BUSCO1668", and 5 subsampled ones, all starting
767 with "T834"; depending on the subsampling strategy, they are identified as "T834 Alignment
768 lengths", "T834 Average bootstrap value", "T834 Completeness", "T834 Treeness / RCV", and
769 T834 Variable sites"). (b) Hierarchical clustering based off of topological similarity values
770 among the 36 phylogenies.
771
772 **Figure 3. The eight internodes not recovered in all 36 phylogenies.**
773 Internode numbers refer to internodes that have at least one conflicting topology among the 36
774 phylogenetic trees inferred from the full and five subsampled data matrices across three different
775 schemes and two data types. The internode recovered from the analysis of the 1,668-gene
776 nucleotide matrix (Figure 1) is shown on the left and the conflicting internode(s) on the right.
777 Next to each of the internodes, the nucleotide (nt) and amino acid (aa) gene support frequency
778 (GSF) values are shown. On the far right, the sequence type, scheme, and data matrix
779 characteristics of the phylogenies that supports the conflicting internodes are shown. Nt and aa
780 sequence types are represented using black and white squares, respectively; partitioned
781 concatenation, non-partitioned concatenation, and coalescence analytical schemes are depicted as
782 black, grey, or white circles, respectively; and the matrix subset is written next to the symbols.
783
784 **Figure 4. The four internodes recovered in all 36 phylogenies but that exhibit very low**
785 **internode certainty values.**
786 Four bipartitions were recovered by all 36 phylogenies but had internode certainty values below
787 0.10. The internode recovered from the analysis of all 36 phylogenies, including of the 1,668-

788 gene nucleotide matrix (Figure 1), is shown on the left and the most prevalent, conflicting
789 internode on the right. Next to each of the internodes, the nucleotide (nt) and amino acid (aa)
790 gene support frequency (GSF) values are shown.
791
792 **Figure 5. A molecular timetree for the family Aspergillaceae.**
793 Blue boxes around each internode correspond to 95% divergence time confidence intervals for
794 each branch of the Aspergillaceae phylogeny. For reference, the geologic time scale is shown
795 right below the phylogeny. Different genera are depicted using different colored boxes;
796 *Aspergillus* is shown in red, *Penicillium* in blue, *Xeromyces* in green, *Monascus* in purple, and
797 *Penicilliopsis* in orange. Different sections within *Aspergillus* and *Penicillium* are depicted with
798 alternating dark grey and grey bars. Dating estimates were calibrated using the following
799 constraints: origin of Aspergillaceae (I2; 50-146 million years ago [mya]), origin of *Aspergillus*
800 (I5; 43-85 mya) the *A. flavus* and *A. oryzae* split (I30; 3.68-3.99 mya), and the *A. fumigatus* and
801 *A. clavatus* split (I38; 35-39 mya); all constraints were obtained from TIMETREE[99].
802

References

1. Houbraken, J., de Vries, R. P. & Samson, R. A. in *Advances in Applied Microbiology* **86,** 199–249 (2014).

2. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. GenBank. *Nucleic Acids Res.* **36,** D25–D30 (2007).

3. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **37,** D5–D15 (2009).

4. Pitt, J. I. in *Advances in experimental medicine and biology* **504,** 29–41 (2002).

5. Ogundero, V. W. Factors affecting growth and cellulose hydrolysis by the thermotolerant aspergillus nidulans from composts. *Acta Biotechnol.* **3,** 65–72 (1983).

6. Gibbons, J. G. & Rokas, A. The function and evolution of the Aspergillus genome. *Trends Microbiol.* **21,** 14–22 (2013).

7. Machida, M., Yamada, O. & Gomi, K. Genomics of Aspergillus oryzae: Learning from the History of Koji Mold and Exploration of Its Future. *DNA Res.* **15,** 173–183 (2008).

8. Gibbons, J. G. *et al.* The Evolutionary Imprint of Domestication on Genome Variation and Function of the Filamentous Fungus Aspergillus oryzae. *Curr. Biol.* **22,** 1403–1409 (2012).

9. Lessard, M.-H., Bélanger, G., St-Gelais, D. & Labrie, S. The Composition of Camembert Cheese-Ripening Cultures Modulates both Mycelial Growth and Appearance. *Appl. Environ. Microbiol.* **78,** 1813–1819 (2012).

10. Nelson, J. H. Production of Blue cheese flavor via submerged fermentation by Penicillium roqueforti. *J. Agric. Food Chem.* **18,** 567–569 (1970).

11. Pel, H. J. *et al.* Genome sequencing and analysis of the versatile cell factory Aspergillus

826 niger CBS 513.88. *Nat. Biotechnol.* **25,** 221–231 (2007).

827 12. Endo, A. A historical perspective on the discovery of statins. *Proc. Japan Acad. Ser. B* **86,**
828 484–493 (2010).

829 13. Hedayati, M. T., Pasqualotto, A. C., Warn, P. A., Bowyer, P. & Denning, D. W.
830 Aspergillus flavus: human pathogen, allergen and mycotoxin producer. *Microbiology* **153,**
831 1677–1692 (2007).

832 14. Nierman, W. C. *et al.* Genomic sequence of the pathogenic and allergenic filamentous
833 fungus Aspergillus fumigatus. *Nature* **438,** 1151–1156 (2005).

834 15. Ballester, A.-R. *et al.* Genome, Transcriptome, and Functional Analyses of Penicillium
835 expansum Provide New Insights Into Secondary Metabolism and Pathogenicity. *Mol.*
836 *Plant-Microbe Interact.* **28,** 232–248 (2015).

837 16. Marcet-Houben, M. *et al.* Genome sequence of the necrotrophic fungus Penicillium
838 digitatum, the main postharvest pathogen of citrus. *BMC Genomics* **13,** 646 (2012).

839 17. Li, B. *et al.* Genomic Characterization Reveals Insights Into Patulin Biosynthesis and
840 Pathogenicity in Penicillium Species. *Mol. Plant-Microbe Interact.* **28,** 635–647 (2015).

841 18. Vining, L. C. Functions of Secondary Metabolites. *Annu. Rev. Microbiol.* **44,** 395–427
842 (1990).

843 19. Keller, N. P., Turner, G. & Bennett, J. W. Fungal secondary metabolism — from
844 biochemistry to genomics. *Nat. Rev. Microbiol.* **3,** 937–947 (2005).

845 20. Macheleidt, J. *et al.* Regulation and Role of Fungal Secondary Metabolites. *Annu. Rev.*
846 *Genet.* **50,** 371–392 (2016).

847 21. Rohlfs, M., Albert, M., Keller, N. P. & Kempken, F. Secondary chemicals protect mould
848 from fungivory. *Biol. Lett.* **3,** 523–525 (2007).

22. Rohlfs, M. & Churchill, A. C. L. Fungal secondary metabolites as modulators of interactions with insects and other arthropods. *Fungal Genet. Biol.* **48,** 23–34 (2011).

23. Squire, R. Ranking animal carcinogens: a proposed regulatory approach. *Science (80-.).* **214,** 877–880 (1981).

24. Fleming, A. On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to Their Use in the Isolation of B. influenzae. *Clin. Infect. Dis.* **2,** 129–139 (1980).

25. Chain, E. *et al.* Penicillin as a chemotherapeutic agent. *Lancet* **236,** 226–228 (1940).

26. Aminov, R. I. A Brief History of the Antibiotic Era: Lessons Learned and Challenges for the Future. *Front. Microbiol.* **1,** (2010).

27. Pitt, J. I. & Taylor, J. W. Aspergillus , its sexual states and the new International Code of Nomenclature. *Mycologia* **106,** 1051–1062 (2014).

28. Samson, R. A. *et al.* Phylogeny, identification and nomenclature of the genus Aspergillus. *Stud. Mycol.* **78,** 141–173 (2014).

29. Salichos, L. & Rokas, A. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* **497,** 327–331 (2013).

30. Rokas, A., Williams, B. L., King, N. & Carroll, S. B. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425,** 798–804 (2003).

31. de Vries, R. P. *et al.* Comparative genomics reveals high biological diversity and specific adaptations in the industrially and medically important fungal genus Aspergillus. *Genome Biol.* **18,** 28 (2017).

32. Nielsen, J. C. *et al.* Global analysis of biosynthetic gene clusters reveals vast potential of secondary metabolite production in Penicillium species. *Nat. Microbiol.* **2,** 17044 (2017).

872 33. Kjærbølling, I. *et al.* Linking secondary metabolites to gene clusters through genome
873 sequencing of six diverse Aspergillus species. *Proc. Natl. Acad. Sci.* **115,** E753–E761
874 (2018).
875 34. Yang, Y. *et al.* Genome Sequencing and Comparative Genomics Analysis Revealed
876 Pathogenic Potential in Penicillium capsulatum as a Novel Fungal Pathogen Belonging to
877 Eurotiales. *Front. Microbiol.* **7,** (2016).
878 35. Hess, J. & Goldman, N. Addressing Inter-Gene Heterogeneity in Maximum Likelihood
879 Phylogenomic Analysis: Yeasts Revisited. *PLoS One* **6,** e22783 (2011).
880 36. Zhong, B., Liu, L., Yan, Z. & Penny, D. Origin of land plants using the multispecies
881 coalescent model. *Trends Plant Sci.* **18,** 492–495 (2013).
882 37. Song, S., Liu, L., Edwards, S. V. & Wu, S. Resolving conflict in eutherian mammal
883 phylogeny using phylogenomics and the multispecies coalescent model. *Proc. Natl. Acad.*
884 *Sci.* **109,** 14942–14947 (2012).
885 38. Shen, X.-X. *et al.* Reconstructing the Backbone of the Saccharomycotina Yeast Phylogeny
886 Using Genome-Scale Data. *G3 Genes|Genomes|Genetics* **6,** 3927–3939 (2016).
887 39. Suh, A. The phylogenomic forest of bird trees contains a hard polytomy at the root of
888 Neoaves. *Zoologica Scripta* **45,** 50–62 (2016).
889 40. Arcila, D. *et al.* Genome-wide interrogation advances resolution of recalcitrant groups in
890 the tree of life. *Nat. Ecol. Evol.* **1,** (2017).
891 41. King, N. & Rokas, A. Embracing Uncertainty in Reconstructing Early Animal Evolution.
892 *Curr. Biol.* **27,** R1081–R1088 (2017).
893 42. Shen, X.-X., Hittinger, C. T. & Rokas, A. Contentious relationships in phylogenomic
894 studies can be driven by a handful of genes. *Nat. Ecol. Evol.* **1,** 0126 (2017).

43. Riley, R. *et al.* Comparative genomics of biotechnologically important yeasts. *Proc. Natl. Acad. Sci.* **113,** 9882–9887 (2016).

44. Kobert, K., Salichos, L., Rokas, A. & Stamatakis, A. Computing the Internode Certainty and Related Measures from Partial Gene Trees. *Mol. Biol. Evol.* **33,** 1606–1617 (2016).

45. Salichos, L., Stamatakis, A. & Rokas, A. Novel Information Theory-Based Measures for Quantifying Incongruence among Phylogenetic Trees. *Mol. Biol. Evol.* **31,** 1261–1271 (2014).

46. Zhou, X. *et al.* in silico Whole Genome Sequencer & Analyzer (iWGS): A Computational Pipeline to Guide the Design and Analysis of de novo Genome Sequencing Studies. *G3 Genes|Genomes|Genetics* (2016). doi:10.1534/g3.116.034249

47. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19,** 455–477 (2012).

48. Chikhi, R. & Medvedev, P. Informed and automated k-mer size selection for genome assembly. *Bioinformatics* **30,** 31–37 (2014).

49. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29,** 1072–1075 (2013).

50. Houbraken, J. & Samson, R. A. Phylogeny of Penicillium and the segregation of Trichocomaceae into three families. *Stud. Mycol.* **70,** 1–51 (2011).

51. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31,** 3210–3212 (2015).

52. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7,** e1002195 (2011).

53. Madden, T. The BLAST sequence analysis tool. *BLAST Seq. Anal. Tool* 1–17 (2013).

54. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 421 (2009).

55. Waterhouse, R. M., Tegenfeldt, F., Li, J., Zdobnov, E. M. & Kriventseva, E. V. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res.* **41,** D358–D365 (2013).

56. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19,** ii215-ii225 (2003).

57. Waterhouse, R. M. *et al.* BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol. Biol. Evol.* **35,** 543–548 (2018).

58. Jarvis, E. D. *et al.* Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science (80-.).* **346,** 1320–1331 (2014).

59. Misof, B. *et al.* Phylogenomics resolves the timing and pattern of insect evolution. *Science (80-.).* **346,** 763–767 (2014).

60. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).

61. Mount, D. W. Using BLOSUM in Sequence Alignments. *Cold Spring Harb. Protoc.* **2008,** pdb.top39-pdb.top39 (2008).

62. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25,** 1422–1423 (2009).

63. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25,** 1972–1973 (2009).

64. Shen, X.-X., Salichos, L. & Rokas, A. A Genome-Scale Investigation of How Sequence,

941 Function, and Tree-Based Gene Properties Influence Phylogenetic Inference. *Genome*
942 *Biol. Evol.* **8,** 2565–2580 (2016).
943 65. Phillips, M. J. & Penny, D. The root of the mammalian tree inferred from whole
944 mitochondrial genomes. *Mol. Phylogenet. Evol.* **28,** 171–185 (2003).
945 66. Felsenstein, J. Evolutionary trees from DNA sequences: A maximum likelihood approach.
946 *J. Mol. Evol.* **17,** 368–376 (1981).
947 67. Edwards, S. V. Is a new and general theory of molecular systematics emerging? *Evolution*
948 *(N. Y).* **63,** 1–19 (2009).
949 68. Mirarab, S. & Warnow, T. ASTRAL-II: coalescent-based species tree estimation with
950 many hundreds of taxa and thousands of genes. *Bioinformatics* **31,** i44–i52 (2015).
951 69. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and
952 Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol.*
953 *Biol. Evol.* **32,** 268–274 (2015).
954 70. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S.
955 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14,**
956 587–589 (2017).
957 71. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
958 large phylogenies. *Bioinformatics* **30,** 1312–1313 (2014).
959 72. Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with
960 variable rates over sites: Approximate methods. *J. Mol. Evol.* **39,** 306–314 (1994).
961 73. Yang, Z. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol.*
962 *Evol.* **11,** 367–372 (1996).
963 74. Tavaré, S. Some probabilistic and statistical problems in the analysis of DNA sequences.

Lect. Math. life Sci. **17,** 57–86 (1986).

75. Vinet, L. & Zhedanov, A. A 'missing' family of classical orthogonal polynomials. *J. Phys. A Math. Theor.* **44,** 085201 (2011).

76. Jones, D. T., Taylor, W. R. & Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Bioinformatics* **8,** 275–282 (1992).

77. Le, S. Q. & Gascuel, O. An Improved General Amino Acid Replacement Matrix. *Mol. Biol. Evol.* **25,** 1307–1320 (2008).

78. Kosiol, C. & Goldman, N. Different Versions of the Dayhoff Rate Matrix. *Mol. Biol. Evol.* **22,** 193–199 (2005).

79. Zhou, X., Shen, X.-X., Hittinger, C. T. & Rokas, A. Evaluating Fast Maximum Likelihood-Based Phylogenetic Programs Using Empirical Phylogenomic Data Sets. *Mol. Biol. Evol.* **35,** 486–503 (2018).

80. Chernomor, O., von Haeseler, A. & Minh, B. Q. Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Syst. Biol.* **65,** 997–1008 (2016).

81. Waddell, P. J. & Steel, M. . General Time-Reversible Distances with Unequal Rates across Sites: Mixing Γ and Inverse Gaussian Distributions with Invariant Sites. *Mol. Phylogenet. Evol.* **8,** 398–414 (1997).

82. Felsenstein, J. The Newick tree format. *English* (1986).

83. Felsenstein, J. Inferring phytogenies. *Sunderland* (2003).

84. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35,** 518–522 (2018).

85. Vinh, L. S. IQPNNI: Moving Fast Through Tree Space and Stopping in Time. *Mol. Biol. Evol.* **21,** 1565–1571 (2004).

987 86. Minh, B. Q., Vinh, L. S., von Haeseler, A. & Schmidt, H. A. pIQPNNI: parallel
988 reconstruction of large maximum likelihood phylogenies. *Bioinformatics* **21,** 3794–3796
989 (2005).
990 87. Kishino, H., Miyata, T. & Hasegawa, M. Maximum likelihood inference of protein
991 phylogeny and the origin of chloroplasts. *J. Mol. Evol.* **31,** 151–160 (1990).
992 88. Hasegawa, M. & Kishino, H. Accuracies of the simple methods for estimating the
993 bootstrap probability of a maximum likelihood tree. *Mol.Biol.Evol.* **11,** 142–145 (1994).
994 89. Stamatakis, A., Hoover, P. & Rougemont, J. A Rapid Bootstrap Algorithm for the
995 RAxML Web Servers. *Syst. Biol.* **57,** 758–771 (2008).
996 90. Felsenstein, J. Confidence Limits on Phylogenies: An Approach Using the Bootstrap.
997 *Evolution (N. Y).* **39,** 783 (1985).
998 91. R Development Core Team, R. *Computational Many-Particle Physics*. *R Foundation for*
999 *Statistical Computing* **739,** (Springer Berlin Heidelberg, 2008).
1000 92. Lê, S., Josse, J. & Husson, F. FactoMineR : An R Package for Multivariate Analysis. *J.*
1001 *Stat. Softw.* **25,** 1–18 (2008).
1002 93. Kassambara, A. & Mundt, F. factoextra. *R package, v. 1.0.5*
1003 http://www.sthda.com/english/rpkgs/factoextra/ (2017).
1004 94. Zhou, X. *et al.* Quartet-based computations of internode certainty provide accurate and
1005 robust measures of phylogenetic incongruence. *bioRxiv* 168526 (2017).
1006 doi:10.1101/168526
1007 95. Junier, T. & Zdobnov, E. M. The Newick utilities: high-throughput phylogenetic tree
1008 processing in the UNIX shell. *Bioinformatics* **26,** 1669–1670 (2010).
1009 96. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **24,**

1586–1591 (2007).

97. DOS REIS, M. & YANG, Z. The unbearable uncertainty of Bayesian divergence time estimation. *J. Syst. Evol.* **51,** 30–43 (2013).

98. Liu, L. *et al.* Genomic evidence reveals a radiation of placental mammals uninterrupted by the KPg boundary. *Proc. Natl. Acad. Sci.* **114,** E7282–E7290 (2017).

99. Hedges, S. B., Dudley, J. & Kumar, S. TimeTree: A public knowledge-base of divergence times among organisms. *Bioinformatics* **22,** 2971–2972 (2006).

100. Berbee, M. L. & Taylor, J. W. in *The Mycota* 229–246 (2001).

101. Vijaykrishna, D., Jeewon, R. & Hyde, K. Molecular taxonomy, origins and evolution of freshwater ascomycetes. *Fungal Divers.* **23,** 351–390 (2006).

102. Sharpton, T. J. *et al.* Comparative genomic analyses of the human fungal pathogens Coccidioides and their relatives. *Genome Res.* **19,** 1722–1731 (2009).

103. Da Lage, J.-L., Binder, M., Hua-Van, A., Janeček, Š. & Casane, D. Gene make-up: rapid and massive intron gains after horizontal transfer of a bacterial α-amylase gene to Basidiomycetes. *BMC Evol. Biol.* **13,** 40 (2013).

104. Kensche, P. R., Oti, M., Dutilh, B. E. & Huynen, M. A. Conservation of divergent transcription in fungi. *Trends Genet.* **24,** 207–211 (2008).

105. Beimforde, C. *et al.* Estimating the Phanerozoic history of the Ascomycota lineages: Combining fossil and molecular data. *Mol. Phylogenet. Evol.* **78,** 386–398 (2014).

106. Fan, H.-W. *et al.* Genomic Analysis of an Ascomycete Fungus from the Rice Planthopper Reveals How It Adapts to an Endosymbiotic Lifestyle. *Genome Biol. Evol.* **7,** 2623–2634 (2015).

107. Gaya, E. *et al.* The adaptive radiation of lichen-forming Teloschistaceae is associated with

1033 sunscreening pigments and a bark-to-rock substrate shift. *Proc. Natl. Acad. Sci.* **112,**
1034 11600–11605 (2015).
1035 108. Raftery, A. E. & Lewis, S. M. The number of iterations, convergence diagnostics and
1036 generic Metropolis algorithms. *Pract. Markov Chain Monte Carlo* **7,** 763–773 (1995).
1037 109. Sedgwick, P. Spearman's rank correlation coefficient. *BMJ* **349,** g7327–g7327 (2014).
1038 110. Harrell Jr, F. E. Package 'Hmisc' (v4.0-0). *URL https://cran.r-*
1039 *project.org/web/packages/Hmisc/index.html* (2015).
1040 111. Wickham, H. *ggplot2. Elegant Graphics for Data Analysis* (Springer New York, 2009).
1041 doi:10.1007/978-0-387-98141-3
1042 112. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of
1043 intersecting sets and their properties. *Bioinformatics* **33,** 2938–2940 (2017).
1044 113. Kolde, R. Package `pheatmap'. *Bioconductor* 1–6 (2012).
1045 114. Rambaut, A. FigTree, a graphical viewer of phylogenetic trees. *Inst. Evol. Biol. Univ.*
1046 *Edinburgh* (2009).
1047 115. Bell, M. A. & Lloyd, G. T. strap: an R package for plotting phylogenies against
1048 stratigraphy and assessing their stratigraphic congruence. *Palaeontology* **58,** 379–389
1049 (2015).
1050 116. Kocsubé, S. *et al.* Aspergillus is monophyletic: Evidence from multiple gene phylogenies
1051 and extrolites profiles. *Stud. Mycol.* **85,** 199–213 (2016).
1052 117. Cary, J. W., Ehrlich, K. C., Beltz, S. B., Harris-Coward, P. & Klich, M. A.
1053 Characterization of the Aspergillus ochraceoroseus aflatoxin/sterigmatocystin biosynthetic
1054 gene cluster. *Mycologia* **101,** 352–362 (2009).
1055 118. Moore, G. G., Mack, B. M. & Beltz, S. B. Draft Genome Sequences of Two Closely

Related Aflatoxigenic Aspergillus Species Obtained from the Ivory Coast. *Genome Biol. Evol.* **8,** 729–732 (2016).

119. Frisvad, J. C., Skouboe, P. & Samson, R. A. Taxonomic comparison of three different groups of aflatoxin producers and a new efficient producer of aflatoxin B1, sterigmatocystin and 3-O-methylsterigmatocystin, Aspergillus rambellii sp. nov. *Syst. Appl. Microbiol.* **28,** 442–453 (2005).

120. Endo, A., Kuroda, M. & Tsujita, Y. ML-236A, ML-236B, and ML-236C, new inhibitors of cholesterogensis produced by Penicillium citrinum. *J. Antibiot. (Tokyo).* **29,** 1346–1348 (1976).

121. Degnan, J. H. & Salter, L. A. Gene tree distributions under the coalescent process. *Evolution (N. Y).* **59,** 24–37 (2005).

122. Sang, T. & Zhong, Y. Testing Hybridization Hypotheses Based on Incongruent Gene Trees. *Syst. Biol.* **49,** 422–434 (2000).

123. Hallett, M., Lagergren, J. & Tofigh, A. Simultaneous identification of duplications and lateral transfers. in *Proceedings of the eighth annual international conference on Computational molecular biology - RECOMB '04* 347–356 (ACM Press, 2004). doi:10.1145/974614.974660

124. Doolittle, W. F. & Bapteste, E. Pattern pluralism and the Tree of Life hypothesis. *Proc. Natl. Acad. Sci.* **104,** 2043–2049 (2007).

125. Castoe, T. A. *et al.* Evidence for an ancient adaptive episode of convergent molecular evolution. *Proc. Natl. Acad. Sci.* **106,** 8986–8991 (2009).

126. Li, Y., Liu, Z., Shi, P. & Zhang, J. The hearing gene Prestin unites echolocating bats and whales. *Curr. Biol.* **20,** R55–R56 (2010).

127. Patterson, N., Richter, D. J., Gnerre, S., Lander, E. S. & Reich, D. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* **441,** 1103–1108 (2006).

128. Hobolth, A., Christensen, O. F., Mailund, T. & Schierup, M. H. Genomic Relationships and Speciation Times of Human, Chimpanzee, and Gorilla Inferred from a Coalescent Hidden Markov Model. *PLoS Genet.* **3,** e7 (2007).

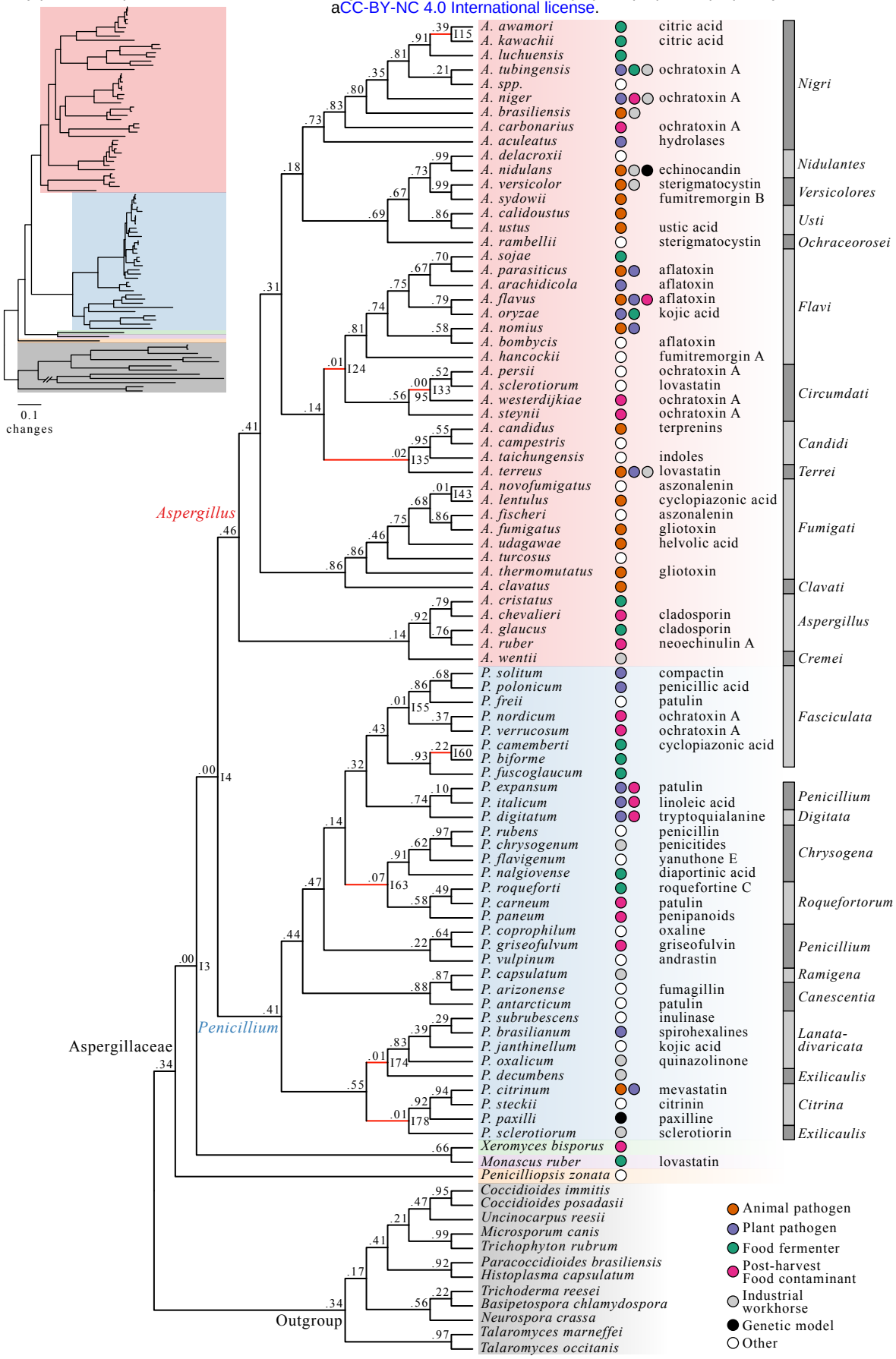
129. Martin, S. H. *et al.* Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Res.* **23,** 1817–1828 (2013).

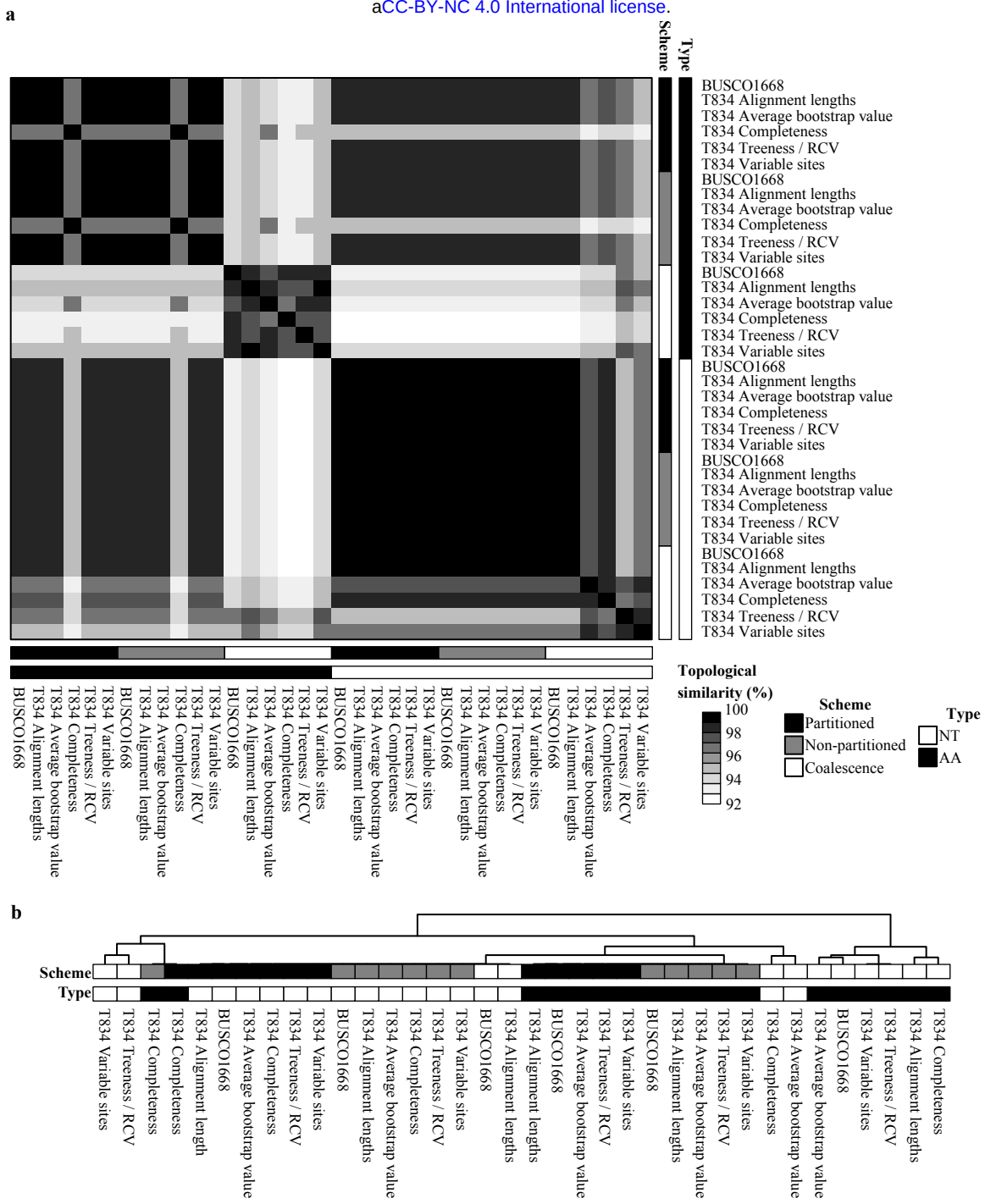
130. Phillips, M. J., Delsuc, F. & Penny, D. Genome-Scale Phylogeny and the Detection of Systematic Biases. *Mol. Biol. Evol.* **21,** 1455–1458 (2004).

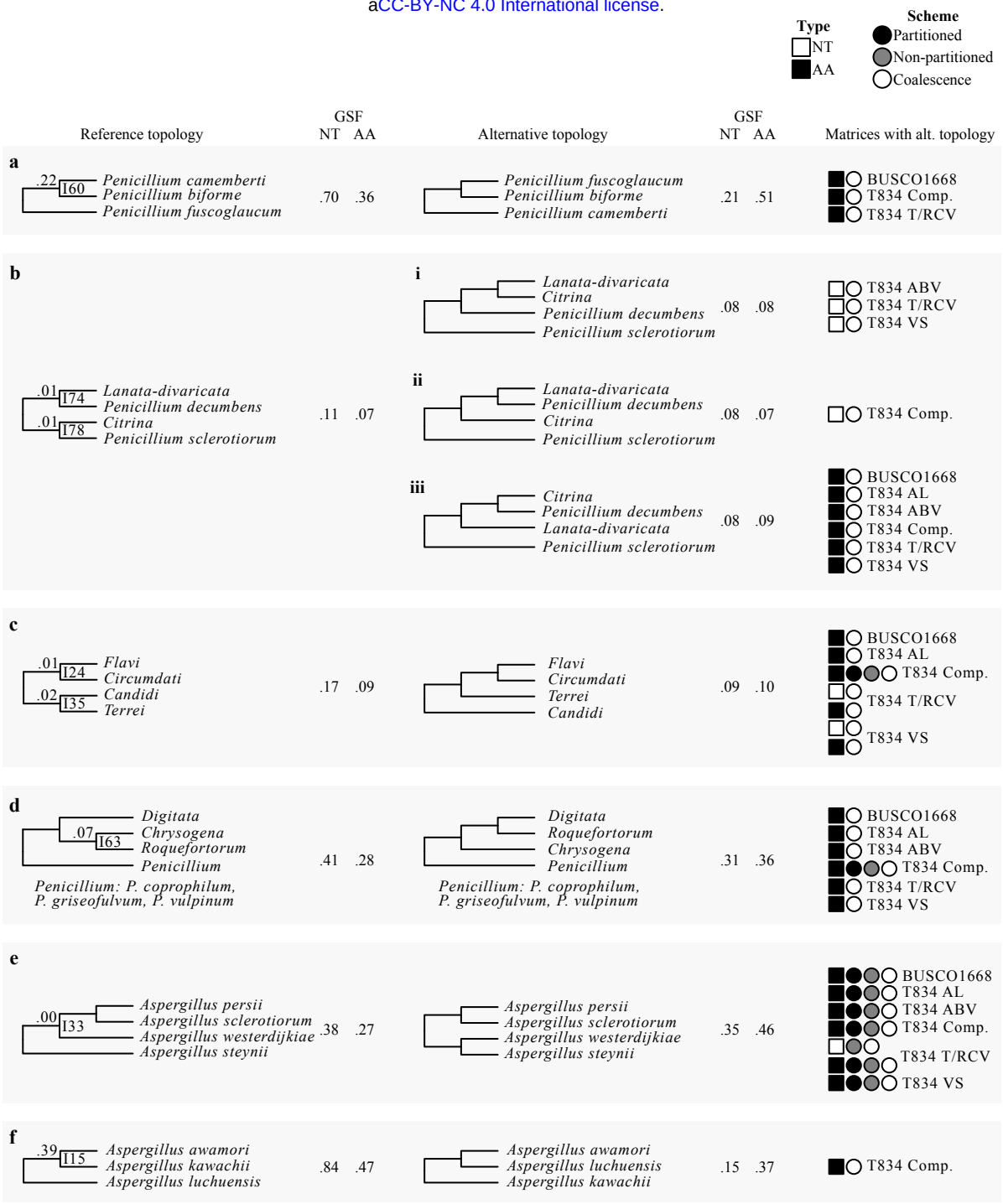
131. Nabhan, A. R. & Sarkar, I. N. The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. *Brief. Bioinform.* **13,** 122–134 (2012).

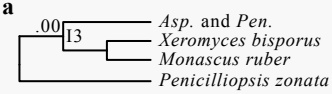
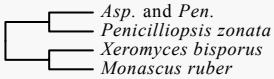
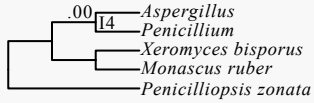
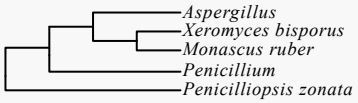
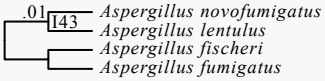
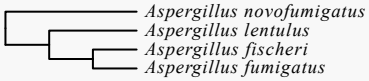
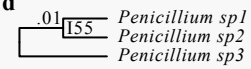
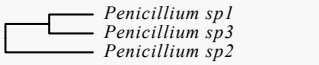
132. Rokas, A. & Carroll, S. B. More Genes or More Taxa? The Relative Contribution of Gene Number and Taxon Number to Phylogenetic Accuracy. *Mol. Biol. Evol.* **22,** 1337–1344 (2005).

133. Baldauf, S. L. Phylogeny for the faint of heart: a tutorial. *Trends Genet.* **19,** 345–351 (2003).







	Reference topology	GSF NT	GSF AA	Alternative topology	GSF NT	GSF AA
a		.12	.15		.08	.12
b		.24	.28		.20	.14
c		.39	.22		.25	.14
d	 *sp1: P. solitum, P. polonicum, P. freii* *sp2: P. nordicum, P. verrucosum* *sp3: P. camemberti, P. biforme, P. fuscoglaucum*	.51	.31	 *sp1: P. solitum, P. polonicum, P. freii* *sp2: P. nordicum, P. verrucosum* *sp3: P. camemberti, P. biforme, P. fuscoglaucum*	.23	.22